多模态内容安全审核系统构建思考

刘 帆 王凤美
（太极计算机股份有限公司，北京 100102）

摘要：【目的】互联网、智能设备及各种新生业务的飞速发展使海量的互联网信息夹杂着大量暴力敏感、低俗等垃圾信息。随着国家对内容安全监管的日渐严格，本文研究实现对海量的互联网信息的快速、精准内容安全审核的方法。【方法】主要运用大数据、人工智能技术对数字内容审核、过滤方式进行革新。【结果】实现将新技术与传统编审机制进行融合。【结论】将人力密集、脑力密集向创新密集、技术密集转型升级，是解决媒体行业跨模态内容安全审核困境的有效途径和必然发展趋势。

关键词：多模态；神经网络；大数据；人工智能；深度学习 **中图分类号：**G642 **文献标识码：**A
文章编号：1671-0134（2023）04-149-05 **DOI：**10.19483/j.cnki.11-4653/n.2023.04.031
本文著录格式：刘帆，王凤美．多模态内容安全审核系统构建思考 [J]. 中国传媒科技，2023（04）：149-153.

导语

在媒体数字化迅速发展的今天，网络信息的发布环境日益复杂，内容良莠不齐。同时，自媒体时代的到来也带来了爆发式增长的内容体量和种类，内容生产、传播的形式发生了巨大变化，传统的内容审核与监管方式耗费的资源在大幅增长，但是工作效率却难以提高。随着互联网、智能设备及各种新生业务的飞速发展，每天通过互联网上传的视频、图片、语音数量超过 10 亿，通过各种社交网络、媒体平台的发文数量超过 5 亿，而且这种趋势还是继续快速增长，但是海量的互联网信息中部分夹杂着暴力敏感、低俗等垃圾信息。[4] 国家针对内容安全监管日渐严格，传统的文字校对及过滤系统已经不能满足快速发展的移动互联网时代内容安全审核的需求，跨模态内容安全审核技术成为互联网不得不面临的严峻挑战。[4]

传统大型网站的内容审核习惯于采用人工审核的形式，审核人员对内容信息逐条进行分析判断，不管是效率还是准确性都难以得到保证。随着人工智能技术的不断成熟，自然语言处理、图像识别、声纹识别等技术已可成熟应用于大部分数字媒体领域。[2] 人工智能领域的深度学习和自然语言处理算法的飞速发展给上述问题带来了创新式的解决方案，不仅能够精准识别内容中出现的风险敏感信息，还能够极大降低内容审核的人力成本。人工智能技术的引入将彻底改变传统的内容审核形式，实现对互联网内容信息的实时审核，无论是审核效率还是审核精度，都将得到极大提升。[1] 运用大数据、人工智能技术对数字内容审核、过滤方式进行革新，将高科技与传统编审机制进行融合，将人力密集、脑力密集向创新密集、技术密集转型升级，是解决媒体行业跨模态内容安全审核困境的有效途径和必然发展趋势。

1. 系统架构设计

多模态内容审核是内容安全审核系统的核心能力，依托行业内领先的深度学习技术、自然语言处理、图像文字识别（OCR）、自动语音识别（ASR）等技术，系统可提供针对图片、视频、文字、音频等多媒体内容风险智能识别服务。在服务模式上灵活多样化，支持 API/SDK、SaaS、私有化部署等多种服务方式，须保证高可用、高性能等特性。[8]

多模态内容安全审核系统是一个复杂的系统，包含诸多模块，这些模块按照功能可以划分为一个典型的层级架构，如图 1：

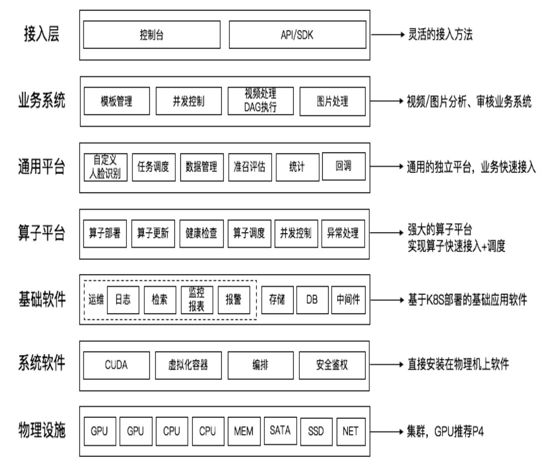


图 1 多模态内容安全审核系统业务架构

（1）物理设施：一组物理机构成的集群，提供了 GPU（推荐 P4）、CPU、内存、磁盘、网卡等物理资源。

（2）系统软件：指直接安装在物理机上的软件。虚拟化与编排利用开源软件 Docker 和 K8S 对物理机集群进行资源虚拟化，并提供编排 API 供上游调度和使用。上层的软件都是通过 K8S 进行安装和部署的。安全鉴权依赖外插式加密狗实现的鉴权服务，保证了多模态内容安全审核系统提供的多个算子，尤其是 AI 算子的安全性。

（3）基础软件：包含了常见的基础应用软件。基于 ELK 搭建的日志收集、检索、查看平台；基于 Prometheus+Grafana 实现的监控项统计、查看、报警平台。基于 Ceph 搭建的分布式文件系统，支持对象存储、块设备存储、文件系统服务。数据库使用高可用 MySQL 集群。中间件为高可用 Redis 集群、ZooKeeper 集群。

（4）算子平台：算子是多模态内容安全审核系统的最小计算单元，算子的输入可以是一个视频，也可以是视频中提取的音频、图片序列，既可能是其他算子的输出，也可能是这几种可能输入的组合。算子平台主要提供算子的自动化运维和算子任务调度两种能力。

（5）通用平台：通用平台不是一个平台，而是多个相对独立的功能模块 / 子系统的合集，这些子系统不直接暴露给用户，但是会被上层的业务系统所依赖。主要包括下列子系统。

● 自定义人脸识别系统：提供了自定义人脸底库管理、人脸特征计算（依赖算子平台）、人脸检索能力。

● 数据管理平台：在视频 AI 一体机中，输入的视频目前仅支持 URL 格式，数据管理平台提供了数据拉取、视频元信息计算、视频转码、缩略图提取、音频提取与 VAD 切割等能力；同时也支持将视频处理的中间结果进行缓存和复用。

● 评估平台：使用一批已标注的数据进行测试与评估是判断一体机效果的常用方法，这一过程往往需要人工进行发起预测、分析对比预测结果与统计计算等繁琐的操作。

● 统计平台：统计平台为一体机提供了服务日志查询与业务信息统计的能力。通过对服务产生日志的收集与存储，提供一个统一的日志与统计数据查询入口，为用户提供问题排查与感知业务变化趋势的途径。

● 回调服务：提供了在特定时间向特定地址发送特定消息的能力，同时支持了简单的重试、并发控制策略。

● 任务调度平台：支撑了视频 AI 一体机中多个模块 / 子系统的调度（例如 算子平台、数据管理平台、评估平台）能力，按照任务队列进行任务调度隔离，实现了队列自动创建和清理、拥塞控制、调度并发配置等高级功能，实现了一个灵活高吞吐的任务调度平台。

（6）业务系统：该系统支撑了多模态内容安全审核系统不同业务场景（如文本分析、图片分析、视频分析）的业务执行。核心的功能包括：

● 模板管理：模板配置了用户期望对视频进行处理的算子类型，分析任务对应不同的模板。

● 并发控制：根据系统集群的大小，业务系统需要控制并发处理的视频数量。

● 视频处理 DAG 执行：根据模板配置，业务系统内部会为每一个视频处理任务生成算子执行路线图（构成一个 DAG），业务系统需要根据该 DAG 依次执行每个算子（调用通用平台及算子平台），最终输出结果。

● 图片处理：支持图片分析、审核等业务。

● 文本处理：支持文本审核等业务。

（7）接入层：两种接入方式

● 控制台：可视化操作界面，可以对模板进行创建、修改、查看，同时也可以发起审核任务并查看结果。

● API/SDK：调用业务系统 Restful HTTP API 进行使用。

2. 审核能力及识别的典型风险场景设计

2.1 文本检测

文本检测基于海量文本特征库、规则库、关键词库、NLP 算法文本进行过滤分析，帮助内容生产者检测制定的文本中是否包含违规信息，例如，对涉黄、涉恐、涉政、广告、违禁、辱骂、低质灌水、负面评论、意识形态风险预警等多种维度进行审核，并支持自定义文本黑库。应支持识别的典型风险场景如表 1。

表 1 文本检测典型风险场景描述表

场景名称	描述
广告	识别检测文本包含电话、微信号、QQ 号、URL、签到、引导签名、搜索等信息
涉政	识别检测文本涉及涉政负面、涉政不确定、人物、人物演绎、事件、事件演绎
辱骂	识别检测文本包括严重、一般、口头语辱骂信息
色情	识别检测文本包括色情违禁、性知识、内涵等内容
兼职代理	识别检测文本包括兼职、上屏、金融短信等内容
自定义	识别检测文本命中自定义关键词

2.2 图片检测

图片检测应用人工智能积极学习算法，通过深度学习模型快速检测出包含色情、涉政、暴恐、垃圾广告、图文违规、图片 Logo 等违规内容。应支持识别的典型风险场景包括表 2 内容：

表 2 图片检测典型风险场景描述表

场景名称	描述	检测结果分类
图片智能鉴黄	检测图片是否包含色情、性感内容	正常、色情、性感
图片暴恐涉政	检测图片是否包含暴恐或涉政类内容	正常、血腥、爆炸烟光、特殊装束、特殊标识、武器、涉政、打斗、聚众、游行、车祸现场、旗帜、地标等
图文违规	检测图片是否包含广告和文字违规信息	正常、含涉政内容、含涉黄内容、含辱骂内容、含暴恐内容、含违禁内容、含其他垃圾内容
图片二维码	检测图片是否含二维码或小程序码	正常、含二维码
图片 logo	检测图片是否包含 logo 信息，例如台标、商标等	正常、受管控的 logo、商标

2.3 音频检测

语音内容审核帮助内容生产者检测音频文件或语音流（例如直播流）中的风险或违规内容，例如垃圾信息、广告、涉政、暴恐、辱骂、色情、灌水、违禁、无意义内容。应支持识别的风险场景见表 3：

表 3 音频检测典型风险场景描述表

场景名称	描述
广告	检测音频包含电话、微信号、QQ 号、URL 等，引导签名、搜索、签到等信息
涉政	检测音频涉及涉政负面、涉政不确定、人物、人物演绎、事件、事件演绎
辱骂	检测音频包括严重、一般、口头语等辱骂信息
色情	检测音频包括色情违禁、性知识、内涵、娇喘呻吟等内容
兼职代理	检测音频包括兼职、上屏、金融短信等内容。
自定义	检测音频命中自定义关键词

2.4 视频检测

视频检测应区分视频文件与直播流，视频文件通过对视频 URL 地址解析下载视频后支持默认时间截帧和用户自顶底截帧频率，进行截帧后图像检测识别。直播则通过拉流的方式，获取视频流数据，并自动将视频转化成图片（按照设定频率截帧），然后对截取的图片进行过滤检测，应支持点播视频、直播视频的过滤。[6] 系统应支持同步与异步两种接口识别接入，异步检测任务不会实施返回检测结果，用户需要通过 Callback 或者轮询的方式获取检测结果。[7–9] 视频检测应支持识别的风险场景见表 4：

表 4 视频检测典型风险场景描述表

场景名称	描述	检测结果分类
视频智能鉴黄	检测视频中是否包含色情内容	正常、色情
视频暴恐涉政	检测视频中是否包含暴恐涉政内容。	正常、暴恐涉政
视频 logo	检测视频中是否包含特定的 logo。	正常、包含 logo
视频图文违规	检测视频中是否包含广告或违规的文字内容。	正常、广告或文字违规
视频语音违规	检测视频中的语音内容是否包含违规信息	正常、含垃圾信息、广告、涉政、暴恐、辱骂、色情、灌水、违禁、自定义（例如命中自定义关键词）

3. 算法识别能力及关键技术介绍

3.1 鉴黄识别

图像鉴黄是利用超大规模数据所提供的知识来对深度神经网络进行引导，训练出多个极具泛化能力的网络模型，同时基于蒸馏学习的思想对模型复杂度和参数规模进行大幅压缩，快速准确地识别出色情、低俗的图片和视频，解决对违规内容识别的问题。

3.1.1 步骤

视频检测首先进行预处理，提取关键图像帧，转换为对图像的检测。

图像通过预先训练的卷积网络提取特征。

提取的特征通过全卷积网络进行二分类，确定是否为色情、低俗图片。

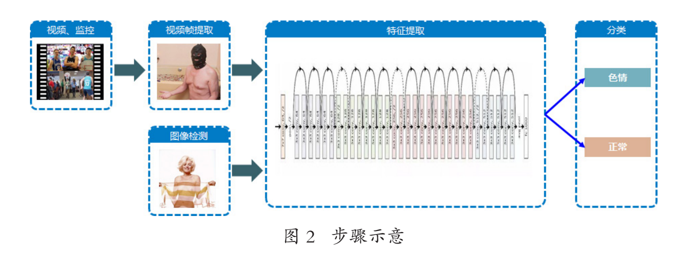


图 2 步骤示意

3.1.2 原理

想要教会机器去识别色情图像，需利用成千上万的图片样本去“训练”它，提取色情图片特征并不断记忆。每张图片中的任何一个点都包括亮度值、色相值以及饱和度值，通过设置这三个值的大小范围，机器能识别出“肉色”，进而猜测出图片里裸露的人体皮肤区域。[7]

色情图片最明显的特点就是画面中人体皮肤颜色所占比例较大，当机器识别图片中有类似人体肤色区域后，需要进一步确认区域的来源，看他们是没有穿衣服的女主角还是正常物体。假设两块黄色区域分别

是两条腿或者两只胳膊，另一块区域是人的身体，这些区域的长度值、宽度值符合人体大小比例，且彼此位置满足一定的几何关系，则有很大可能是色情图片，如果这些区域之间大小和位置不像是人的身体，则可以排除色情图片的嫌疑。

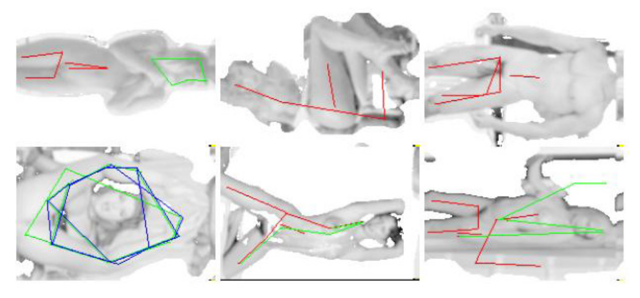


图 3　计算肤色区域的几何关系

3.1.3　分类标准

色情：裸露敏感部位，包含露骨镜头，描述性交行为和色情场景的图片；

性感：衣着暴露但没有裸露敏感部位；

正常：非色情，非性感图片。

3.2　鉴黄识别

图像暴恐识别是通过海量暴恐图片和视频数据源，依托分布式深度学习平台，准确地对图片视频进行暴恐分类，具体支持血腥类、爆炸类、斩首、游行集会、打架斗殴、警民冲突、恐怖主义、战争军队、枪支刀具，敏感着装、敏感文字、各种旗帜等不同类型。

3.2.1　步骤

视频首先进行预处理，截取短视频段和关键图像帧。

通过“卷积神经网络”和“循环神经网络” 提取短视频的特征；通过卷积网络提取视频帧特征。

将视频特征和图像特征进行融合。

通过全卷积网络和 softmax 分类函数确定视频的类别。

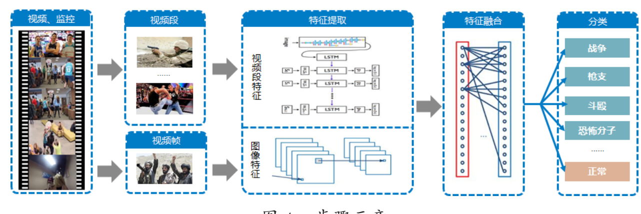


图 4　步骤示意

3.2.2　原理

想要教会机器去识别暴恐图像，同样需利用成千上万的图片样本去“训练”它，提取暴恐图片特征并不断记忆。每一类暴恐图片都有明显的特征标识，例如枪支、匕首、刀具的轮廓，旗帜图案的轮廓，爆炸场景的色差等，通过不断训练机器去记住这些暴恐的特征，从而在新发现图像时快速比对特征值，从而发现暴恐图像。

3.2.3　分类标准

正常：图片中不具备有暴恐特征的图片；

武器 / 持武器者：图片中出现枪支、管制刀具或者其持有者；

特定人物：图片中出现已知恐怖分子头目、政治敏感人物；

特殊符号：图片中（包括书籍）出现的特殊文字；暴恐犯罪组织的标志，部分犯罪分子电视台台标，部分宗教符号； 暴恐反动组织的旗帜；

特殊着装人物：图片中出现穿着迷彩服、军装（包括警察，特警和武警）、特殊服饰等特征；

国家标识：图片中包含有某一个国家的国旗、国徽或者两者；

血腥场景：图片中出现有流血、手术、车祸流血等场景的；

暴乱场景：图中出现有游行、斗殴、焚烧等场景的；

战争场景：图中出现有大型作战武器（如坦克、战斗机）、爆炸、成群军人的。

3.3　政治人物识别

政治人物识别是基于海量人脸库和专业审核人员的审核标准，利用分布式深度学习平台，识别正常、漫画、负面涉政人物的违规信息，降低违规风险，覆盖涉政人物，具体包括国内外国家元首、副国级以上领导人、落马官员、反华势力和劣迹艺人等。

3.4　图文垃圾广告识别

采用深度学习算法，结合图文 OCR 技术、NLP 自然语言处理技术对图片中的图像、文字、水印进行识别，准确识别出含有二维码、垃圾广告、色情、涉政、辱骂等垃圾内容。

垃圾广告：含有大量招嫖、广告、涉黄、辱骂等文字信息的图片。

二维码广告：含有印有二维码、小程序码等内容的图片。

结语

随着媒体融合向纵深推进，5G 技术的加快布局以及大数据、云计算、物联网、区块链、人工智能等多种新兴技术的叠加，中央、各省主流媒体和市县级媒体以先进技术为核心动力引领驱动融合发展，着力向智慧融媒体建设转型，从而重塑了传媒行业的生态格局，如何让融媒体作品智能高效生产的同时保证内容

安全，成为国内外媒体单位、科研机构和技术厂商共同探讨的话题。[5] 目前，国内已有优质人工智能技术厂商，凭借多年产品、技术沉淀，正在加强研究进一步尝试如何更好地为媒体单位提供更加可靠的多模态内容安全审核产品与服务。

参考文献

[1] 王亚辉，王晶 . 人工智能之于科技期刊出版业态的变革及启示 [J]. 中国传媒科技，2023（1）：52−55.

[2] 郭宇辉 . 虚假新闻核查评级机制研究——以 NewsGuard 为例 [J]. 中国传媒科技，2022，（12）：29−32.

[3] 强艳丽 . 新技术对媒体业态的影响及媒体数字化转型研究 [J]. 中国传媒科技，2022（2）：103−105.

[4] 王正芳，赵磊 . 重构传播价值生态链 实现全媒体融合发展——关于广播电视台运营发展的若干思考 [J]. 传播力研究，2018（29）.

[5] 喻国明，刘旸 . 媒介融合时代基于大数据的传媒生产创新 [J]. 传媒观察，2015（9）.

[6] 王慧 . 人工智能技术对播音主持行业的影响与改变 [J]. 传媒论坛，2019（9）：120.

[7] 张皓，吴建鑫 . 基于深度特征的无监督图像检索研究综述 [J]. 计算机研究与发展，2018（9）：1829−1842.

[8] 发明专利《一种视频内容审核系统与方法》[P]. 专利号：CN200610167182.7.

[9] 宋卿，戚成琳，张鹏洲 . 知识图谱技术在新闻领域中的应用思考 [J]. 中国传媒科技，2016（5）：19−21.

作者简介：刘帆（1979−），男，安徽宿州，正高级职称，太极计算机股份有限公司助理总裁，研究方向为媒体融合；王凤美（1988−），女，山东，中级职称，太极计算机股份有限公司，研究方向为媒体融合、媒体大数据应用。

（责任编辑：赵国旭）

（上接第148页）

[2] 宋红波，王雪利 . 近十年国内语料库语言学研究综述 [J]. 山东外语教学，2013（3）：41−47.

[3] 中国人工智能产业发展联盟 . AI 赋能：驱动产业变革的人工智能应用 [M]. 北京： 人民邮电出版社，2019.

[4] 中国人工智能产业发展联盟 . 人工智能浪潮：科技改变生活的 100 个前沿 AI 应用 [M]. 北京： 人民邮电出版社，2018.

[5] 李苏，曾然然，殷治纲 . AI 智能语音技术与产业创新实践 [M]. 北京： 人民邮电出版社，2021.

[6] 荀恩东 . 自然语言结构计算——GPF 结构分析框架 [M]. 北京： 人民邮电出版社，2021.

[7] 荀恩东 . 自然语言结构计算——BCC 语料库 [M]. 北京： 人民邮电出版社，2023.

[8] 荀恩东 . 自然语言结构计算——意合图理论与技术 [M]. 北京： 人民邮电出版社，2023.

[9] 人民日报 . 聚焦 2022 世界人工智能大会，加速赋能实体经济 [EB/OL].https：//sdxw.iqilu.com/share/YS0yMS0xMzM0Mjc3Mw==.html，2022−09−06/2023−03−02.

[10] 人民日报 . 习近平总书记指出，“把新一代人工智能作为推动科技跨越发展、产业优化升级、生产力整体跃升的驱动力量，努力实现高质量发展”[EB/OL].http：//www.mva.gov.cn/sy/xx/szyw/202209/t20220905_65267.html，2022−08−05/2023−03−03.

作者简介：刘亚珍（1987−），女，陕西，责任编辑、策划编辑，研究方向为信息通信类图书出版。

（责任编辑：张晓婧）